Bayesian Reasoning

Chapter 13

Thomas Bayes, 1701-1761

Today's topics

- Review probability theory
- Bayesian inference
  - From the joint distribution
  - Using independence/factoring
  - From sources of evidence
- Naïve Bayes algorithm for inference and classification tasks

Many Sources of Uncertainty

- Uncertain **inputs --** missing and/or noisy data
- Uncertain **knowledge**
  - Multiple causes lead to multiple effects
  - Incomplete enumeration of conditions or effects
  - Incomplete knowledge of causality in the domain
  - Probabilistic/stochastic effects
- Uncertain **outputs**
  - Abduction and induction are inherently uncertain
  - Default reasoning, even deductive, is uncertain
  - Incomplete deductive inference may be uncertain

▸ Probabilistic reasoning only gives probabilistic results

Decision making with uncertainty

Rational behavior:

- For each possible action, identify the possible outcomes
- Compute the **probability** of each outcome
- Compute the **utility** of each outcome
- Compute the probability-weighted **(expected) utility** over possible outcomes for each action
- Select action with the highest expected utility (principle of **Maximum Expected Utility**)

Consider

- Your house has an alarm system
- It should go off if a burglar breaks into the house
- It can go off if there is an earthquake
- How can we predict what's happened if the alarm goes off?

Probability theory 101

- **Random variables**
  - Domain
- **Atomic event**: complete specification of state
- **Prior probability**: degree of belief without any other evidence or info
- **Joint probability**: matrix of combined probabilities of set of variables

- Alarm, Burglary, Earthquake
  - Boolean (like these), discrete, continuous
- Alarm=T∧Burglary=T∧Earthquake=F
  alarm ∧ burglary ∧ ¬earthquake
- P(Burglary) = 0.1
  P(Alarm) = 0.1
  P(earthquake) = 0.000003
- P(Alarm, Burglary) =

	alarm	¬alarm
burglary	.09	.01
¬burglary	.1	.8

Probability theory 101

	alarm	¬alarm
burglary	.09	.01
¬burglary	.1	.8

- **Conditional probability**: prob. of effect given causes
- **Computing conditional probs**:
  - $P(a \mid b) = P(a \wedge b) / P(b)$
  - $P(b)$: **normalizing** constant
- **Product rule**:
  - $P(a \wedge b) = P(a \mid b) * P(b)$
- **Marginalizing**:
  - $P(B) = \Sigma_a P(B, a)$
  - $P(B) = \Sigma_a P(B \mid a)\, P(a)$ (**conditioning**)

- P(burglary | alarm) = .47
  P(alarm | burglary) = .9
- P(burglary | alarm) =
  P(burglary ∧ alarm) / P(alarm)
  = .09/.19 = .47
- P(burglary ∧ alarm) =
  P(burglary | alarm) * P(alarm)
  = .47 * .19 = .09
- P(alarm) =
  P(alarm ∧ burglary) +
  P(alarm ∧ ¬burglary)
  = .09+.1 = .19

Example: Inference from the joint

	alarm		¬alarm	
	earthquake	¬earthquake	earthquake	¬earthquake
burglary	.01	.08	.001	.009
¬burglary	.01	.09	.01	.79

P(burglary | alarm) = α P(burglary, alarm)
= α [P(burglary, alarm, earthquake) + P(burglary, alarm, ¬earthquake)
= α [(.01, .01) + (.08, .09)]
= α [(.09, .1)]

Since P(burglary | alarm) + P(¬burglary | alarm) = 1, α = 1/(.09+.1) = 5.26
(i.e., P(alarm) = 1/α = .19 – **quizlet**: how can you verify this?)

P(burglary | alarm) = .09 * 5.26 = .474

P(¬burglary | alarm) = .1 * 5.26 = .526

Consider

- A student has to take an exam
- She might be smart
- She might have studied
- She may be prepared for the exam
- How are these related?

Exercise:
Inference from the joint

p(smart ∧ study ∧ prep)	smart		¬smart	
	study	¬study	study	¬study
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

Queries:

- **What is the prior probability of *smart*?**
- What is the prior probability of *study*?
- What is the conditional probability of *prepared*, given *study* and *smart*?

Exercise:
Inference from the joint

p(smart ∧ study ∧ prep)	smart		¬smart	
	study	¬study	study	¬study
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

Queries:

- **What is the prior probability of *smart*?**
- What is the prior probability of *study*?
- What is the conditional probability of *prepared*, given *study* and *smart*?

p(smart) = .432 + .16 + .048 + .16 = **0.8**

Exercise: Inference from the joint

p(smart ∧ study ∧ prep)	smart		¬smart	
	study	**¬study**	**study**	**¬study**
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

Queries:

- What is the prior probability of *smart*?
- **What is the prior probability of *study*?**
- What is the conditional probability of *prepared*, given *study* and *smart*?

Exercise:
Inference from the joint

p(smart ∧ study ∧ prep)	smart		¬smart	
	study	¬study	study	¬study
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

Queries:

- What is the prior probability of *smart*?
- **What is the prior probability of *study*?**
- What is the conditional probability of *prepared*, given *study* and *smart*?

p(study) = .432 + .048 + .084 + .036 = **0.6**

Exercise:
Inference from the joint

p(smart ∧ study ∧ prep)	smart		¬smart	
	study	¬study	study	¬study
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

Queries:

- What is the prior probability of *smart*?
- What is the prior probability of *study*?
- **What is the conditional probability of *prepared*, given *study* and *smart*?**

Exercise:
Inference from the joint

p(smart ∧ study ∧ prep)	smart		¬smart	
	study	¬study	study	¬study
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

Queries:

– What is the prior probability of *smart*?

– What is the prior probability of *study*?

– **What is the conditional probability of *prepared*, given *study* and *smart*?**

p(prepared|smart,study)= p(prepared,smart,study)/p(smart, study)
= .432 / (.432 + .048)
= **0.9**

Independence

- When variables don't affect each others' probabilities, we call them **independent**, and can easily compute their joint and conditional probability:

  Independent(A, B) $\rightarrow$ $P(A \wedge B) = P(A) * P(B)$ or $P(A|B) = P(A)$

- {moonPhase, lightLevel} *might* be independent of {burglary, alarm, earthquake}
  - Maybe not: burglars may be more active during a new moon because darkness hides their activity
  - But if we know light level, moon phase doesn't affect whether we are burglarized
  - If burglarized, light level doesn't affect if alarm goes off
- Need a more complex notion of independence and methods for reasoning about the relationships

Exercise: Independence

p(smart ∧ study ∧ prep)	smart		¬smart	
	study	¬study	study	¬study
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

Queries:

- **Q1: Is *smart* independent of *study*?**
- **Q2: Is *prepared* independent of *study*?**

How can we tell?

Exercise: Independence

p(smart ∧ study ∧ prep)	smart		¬smart	
	study	¬study	study	¬study
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

Q1: Is *smart* independent of *study*?

- You might have some intuitive beliefs based on your experience
- You can also check the data

Which way to answer this is better?

Exercise: Independence

p(smart ∧ study ∧ prep)	smart		¬smart	
	study	¬study	study	¬study
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

Q1: Is *smart* independent of *study*?

Q1 true iff p(smart|study) == p(smart)

p(smart|study) = p(smart,study)/p(study)
= (.432 + .048) / .6 = 0.8
0.8 == 0.8, so smart is independent of study

Exercise: Independence

p(smart ∧ study ∧ prep)	smart		¬smart	
	study	¬study	study	¬study
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

Q2: Is *prepared* independent of *study*?

- What is prepared?
- Q2 true iff

Exercise: Independence

p(smart ∧ study ∧ prep)	smart		¬smart	
	study	¬study	study	¬study
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

Q2: Is *prepared* independent of *study*?

Q2 true iff p(prepared|study) == p(prepared)

p(prepared|study) = p(prepared,study)/p(study)

= (.432 + .084) / .6 = .86

0.86 ≠ 0.8, so prepared not independent of study

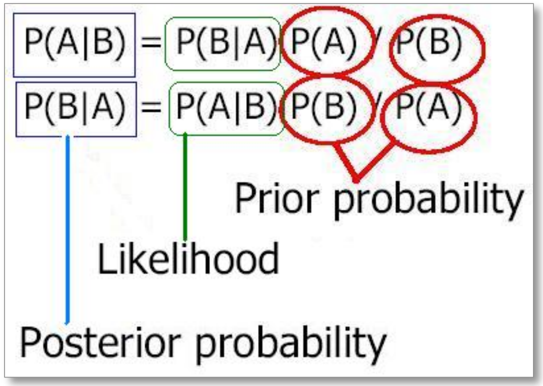
Conditional independence

- Absolute independence:
  - A and B are **independent** if $P(A \wedge B) = P(A) * P(B)$; equivalently, $P(A) = P(A \mid B)$ and $P(B) = P(B \mid A)$
- A and B are **conditionally independent** given C if
  - $P(A \wedge B \mid C) = P(A \mid C) * P(B \mid C)$
- This lets us decompose the joint distribution:
  - $P(A \wedge B \wedge C) = P(A \mid C) * P(B \mid C) * P(C)$
- Moon-Phase and Burglary are ***conditionally independent given*** Light-Level
- Conditional independence is weaker than absolute independence, but useful in decomposing full joint probability distribution

Conditional independence

- Intuitive understanding: conditional independence often arises due to causal relations
  - Moon phase causally effects light level at night
  - Other things do too, e.g., street lights
- For our burglary scenario, moon phase doesn't effect anything else
- Knowing light level means we can ignore moon phase in predicting whether or not alarm suggests we had a burglary

Bayes' rule



Derived from the product rule:

C is a cause, E is an effect

– P(C, E) = P(C|E) * P(E) *# from definition of conditional probability*

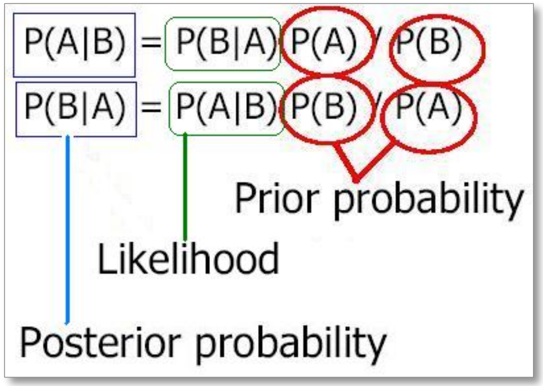
– P(E, C) = P(E|C) * P(C) *# from definition of conditional probability*

– P(C, E) = P(E, C) *# since order is not important*

So…

P(C|E) = P(E|C) * P(C) / P(E)

Bayes' rule



- Derived from the product rule:
  - P(C|E) = P(E|C) * P(C) / P(E)
- **Useful for diagnosis**:
  - If E are (observed) effects and C are (hidden) causes,
  - Often have model for how causes lead to effects P(E|C)
  - May also have prior beliefs (based on experience) about frequency of causes (P(C))
  - Which allows us to reason abductively from effects to causes (P(C|E))

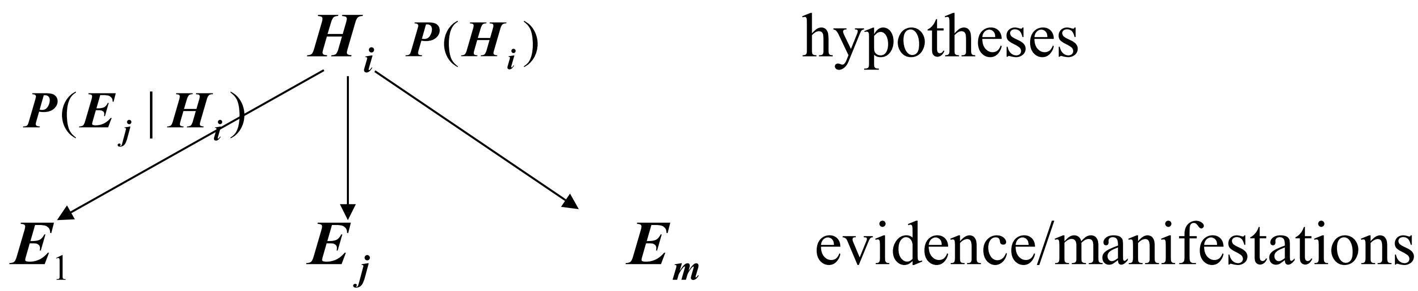
Ex: meningitis and stiff neck

- Meningitis (M) can cause stiff neck (S), though there are other causes too
- Use S as a diagnostic symptom and estimate **p(M|S)**
- Studies can estimate p(M), p(S) & p(S|M), e.g. p(M)=0.7, p(S)=0.01, p(M)=0.00002
- Harder to directly gather data on p(M|S)
- Applying Bayes' Rule:

  $$p(M|S) = p(S|M) * p(M) / p(S) = 0.0014$$

Bayesian inference

- In the setting of diagnostic/evidential reasoning



$P(H_i)$

  – Know prior probability of hypothesis $P(E_j \mid H_i)$

  conditional probability $P(H_i \mid E_j)$

  – Want to compute the *posterior probability*

- Bayes' s theorem:

$$P(H_i \mid E_j) = P(H_i) * P(E_j \mid H_i) / P(E_j)$$

Simple Bayesian diagnostic reasoning

- AKA Naive Bayes classifier
- Knowledge base:
  - Evidence / manifestations: $E_1, \dots E_m$
  - Hypotheses / disorders: $H_1, \dots H_n$

    Note: E_j and H_i are **binary**; hypotheses are **mutually exclusive** (non-overlapping) and **exhaustive** (cover all possible cases)
  - Conditional probabilities: $P(E_j \mid H_i)$, $i = 1, \dots n$; $j = 1, \dots m$
- Cases (evidence for a particular instance): $E_1, \dots, E_l$
- Goal: Find the hypothesis H_i with highest posterior
  - $\text{Max}_i \, P(H_i \mid E_1, \dots, E_l)$

Simple Bayesian diagnostic reasoning

- Bayes' rule says that

  $P(H_i \mid E_1 \ldots E_m) = P(E_1 \ldots E_m \mid H_i)\, P(H_i) / P(E_1 \ldots E_m)$

- Assume each evidence E_i is conditionally independent of the others, *given* a hypothesis H_i, then:

  $P(E_1 \ldots E_m \mid H_i) = \prod^{m}_{j=1} P(E_j \mid H_i)$

- If we only care about relative probabilities for the H_i, then we have:

  $P(H_i \mid E_1 \ldots E_m) = \alpha\, P(H_i) \prod^{m}_{j=1} P(E_j \mid H_i)$

Limitations

- Can't easily handle **multi-fault situations** or cases where intermediate (hidden) causes exist:
  - Disease D causes syndrome S, which causes correlated manifestations M_1 and M_2
- Consider composite hypothesis $H_1 \wedge H_2$, where H_1 & H_2 independent. What's relative posterior?

$$
\begin{aligned}
P(H_1 \wedge H_2 \mid E_1, \ldots, E_l) &= \alpha\, P(E_1, \ldots, E_l \mid H_1 \wedge H_2)\, P(H_1 \wedge H_2) \\
&= \alpha\, P(E_1, \ldots, E_l \mid H_1 \wedge H_2)\, P(H_1)\, P(H_2) \\
&= \alpha \prod_{j=1}^{l} P(E_j \mid H_1 \wedge H_2)\, P(H_1)\, P(H_2)
\end{aligned}
$$

- How do we compute $P(E_j \mid H_1 \wedge H_2)$?

Limitations

- Assume H1 and H2 are independent, given E1, …, El?
  - $P(H_1 \wedge H_2 \mid E_1, \ldots, E_l) = P(H_1 \mid E_1, \ldots, E_l)\, P(H_2 \mid E_1, \ldots, E_l)$
- Unreasonable assumption
  - Earthquake & Burglar independent, but *not* given Alarm:

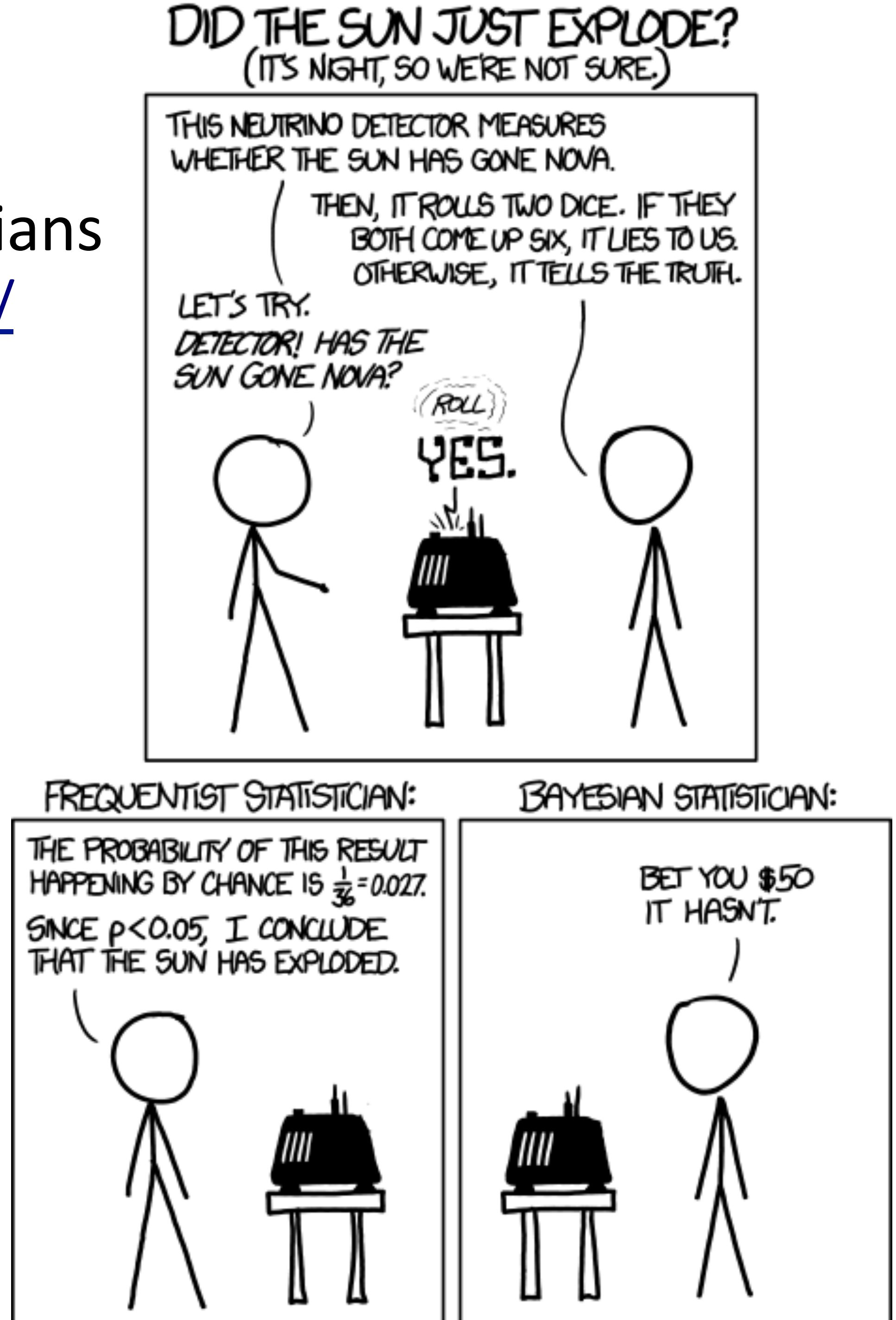
    P(burglar | alarm, earthquake) << P(burglar | alarm)
- Doesn't allow causal chaining:
  - A: 2017 weather; B: 2017 corn production; C: 2018 corn price
  - A influences C indirectly: A→ B → C
  - P(C | B, A) = P(C | B)
- Need richer representation for interacting hypoth- eses, conditional independence & causal chaining
- Next: Bayesian Belief networks!

Summary

- Probability is a rigorous formalism for uncertain knowledge
- Joint probability distribution specifies probability of every atomic event
- Can answer queries by summing over atomic events
- But we must find a way to reduce joint size for non-trivial domains
- Bayes rule lets us compute from known conditional probabilities, usually in causal direction
- Independence & conditional independence provide tools
- Next: Bayesian belief networks

Frequentists vs. Bayesians

http://xkcd.com/1132/



Postscript: Frequentists vs. Bayesians

- **Frequentist inference** draws conclusions from sample data based on frequency or proportion of data
- **Bayesian inference** uses Bayes' rule to update probability estimates for hypothesis as additional evidence is learned
- Differences are often subtle, but can be consequential